REPORT

HANSHIN UNIVERSITY

과목명 | 빅데이터분석실습

담당교수 | 이승천 교수님

학과 | 응용통계학과

학년 | 4학년

학번 | 201452024

이름 | 박상희

제출일 | 2019. 04. 08

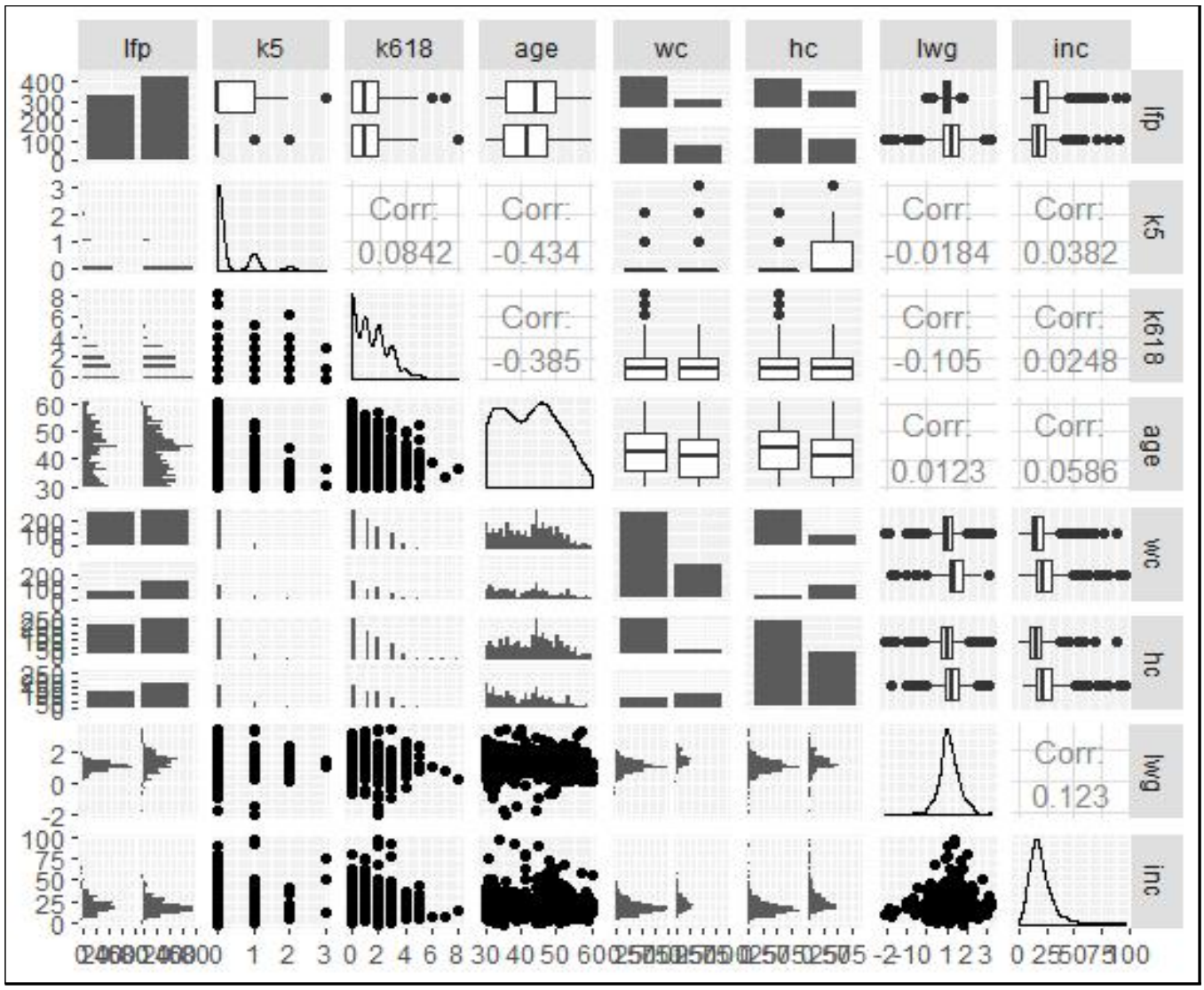
Q. Mroz 데이터에 대해 로지스틱 회귀모형을 적합 시키고, 오분류표를 구해 볼 것.

- 753개의 관찰값 중 임의로 선택된 500개의 데이터로 아래와 같은 로지스틱 회귀모형을 적합하고,
- 오분류표를 구하라.
- $x'\beta = \beta_0 + \beta_1 inc + \beta_2 wc + \beta_3 lwg + \beta_4 lwg^2 + \beta_5 age + \beta_6 k5 + \beta_7 k618$
- $\Pr[lfp_i = yes|x_i] = \dfrac{1}{1+e^{-x_i\beta}}$
- 구해진 모형을 이용하여 나머지 238개의 관찰값에서 오분류표를 구하라.

==========

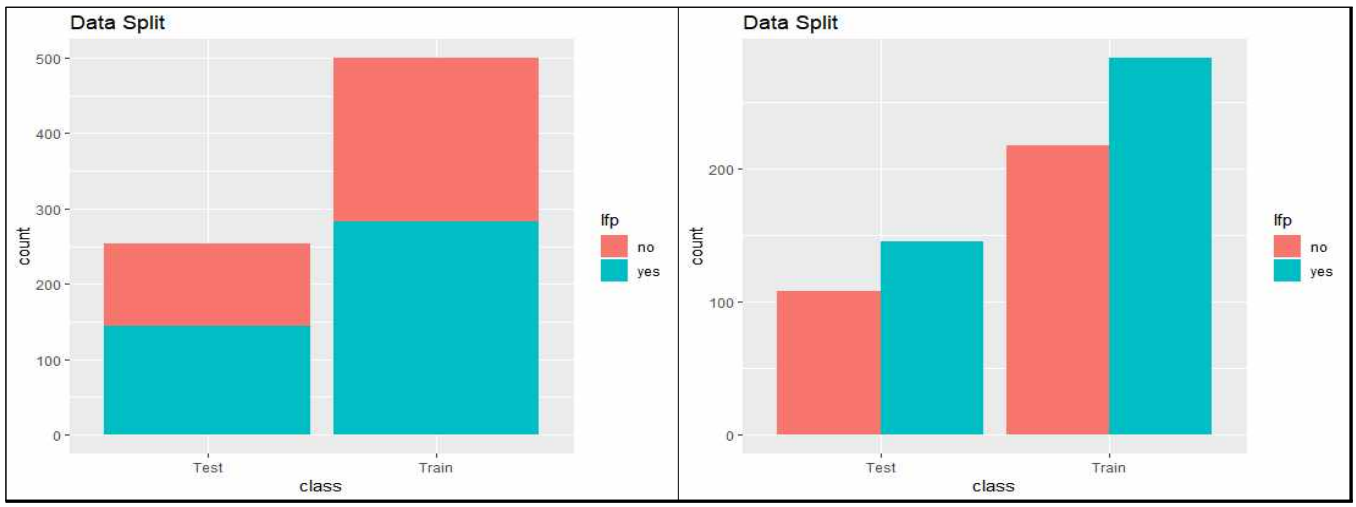
01. 데이터 설명

- 753 obs, 8 variables
- lfp : 결혼한 백인 여성의 직업 여부 (factor)
- inc : 여성을 제외한 가구 수입 (numeric)
- wc : 여성의 대학 졸업 여부 (factor)
- hc : 남편의 대학 졸업 여부 (factor)
- lwg : 여성의 예상 임금 (integer)
- age : 나이 (integer)
- k5 : 5세 이하의 자녀의 수 (integer)
- k618 : 6~18세 이하의 자녀의 수 (integer)



02. 데이터 분리

- Train Data : 500 obs (yes : 217, no : 283)
- Test Data : 253 obs (yes : 108, no : 145)



03. 모형 적합

Summary of the Logistic Regression model (built using glm):

Call:
glm(formula = lfp ~ ., family = binomial(link = "logit"), data = crs$dataset[,
    c(crs$input, crs$target)])

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.19461	-0.88995	0.06788	0.83896	2.25246

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	6.40107	1.08883	5.879	4.13e-09	***
k5	-1.50994	0.28470	-5.304	1.14e-07	***
k618	-0.07657	0.09460	-0.809	0.418249	
age	-0.07152	0.01766	-4.050	5.12e-05	***
wcyes	0.13742	0.30243	0.454	0.649546	
lwg	-6.80843	1.21875	-5.586	2.32e-08	***
inc	-0.04485	0.01179	-3.804	0.000142	***
lwg2	4.20563	0.65194	6.451	1.11e-10	***

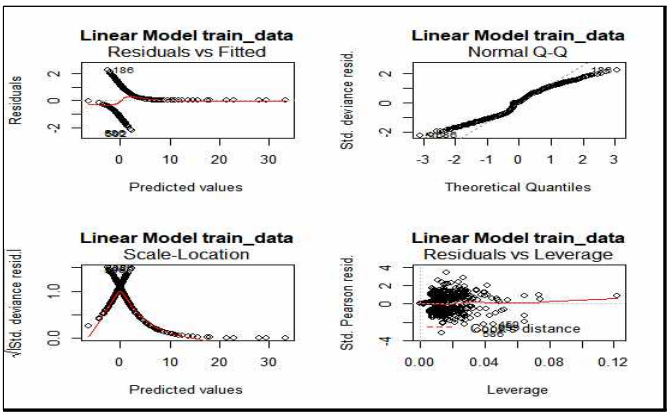
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 684.41 on 499 degrees of freedom
Residual deviance: 491.38 on 492 degrees of freedom
AIC: 507.38

Number of Fisher Scoring iterations: 7

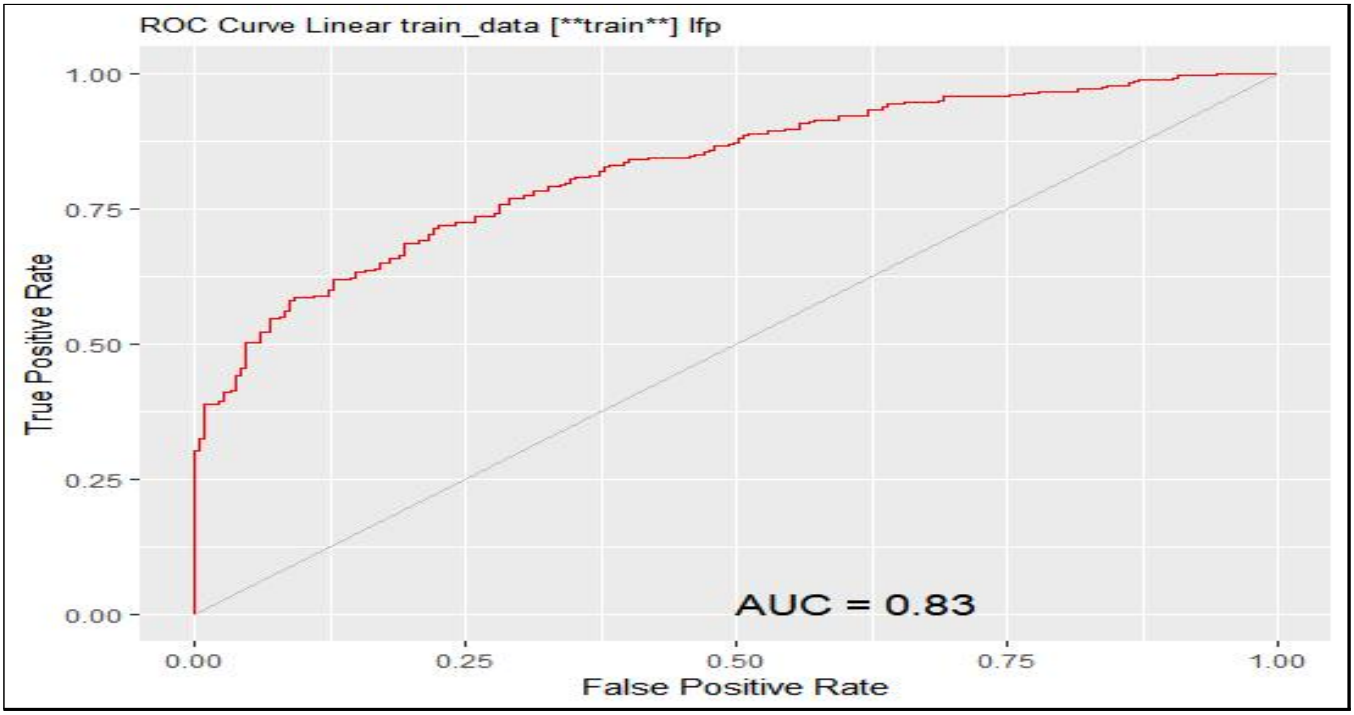
Log likelihood: -245.690 (8 df)
Null/Residual deviance difference: 193.030 (7 df)
Chi-square p-value: 0.00000000
Pseudo R-Square (optimistic): 0.56306669



04. 모델평가

Train Data (500 obs)		Predicted	
		no	yes
Actual	no	154	63
	yes	68	215

특이도(Specificity)	$\frac{154}{63+154} = 0.71$
민감도(Sensitivity)	$\frac{215}{68+215} = 0.77$



05. Test Data

Test Data (253 obs)		Predicted	
		no	yes
Actual	no	78	30
	yes	35	110

특이도(Specificity)	$\frac{78}{78+30} = 0.72$
민감도(Sensitivity)	$\frac{110}{110+35} = 0.76$